データマイニングと情報可視化

Week 4

稲垣 紫緒
いながき　しお

理学研究院　物理学部門 / 共創学部
inagaki@phys.kyushu-u.ac.jp
ウェスト1号館 W1-A823号室

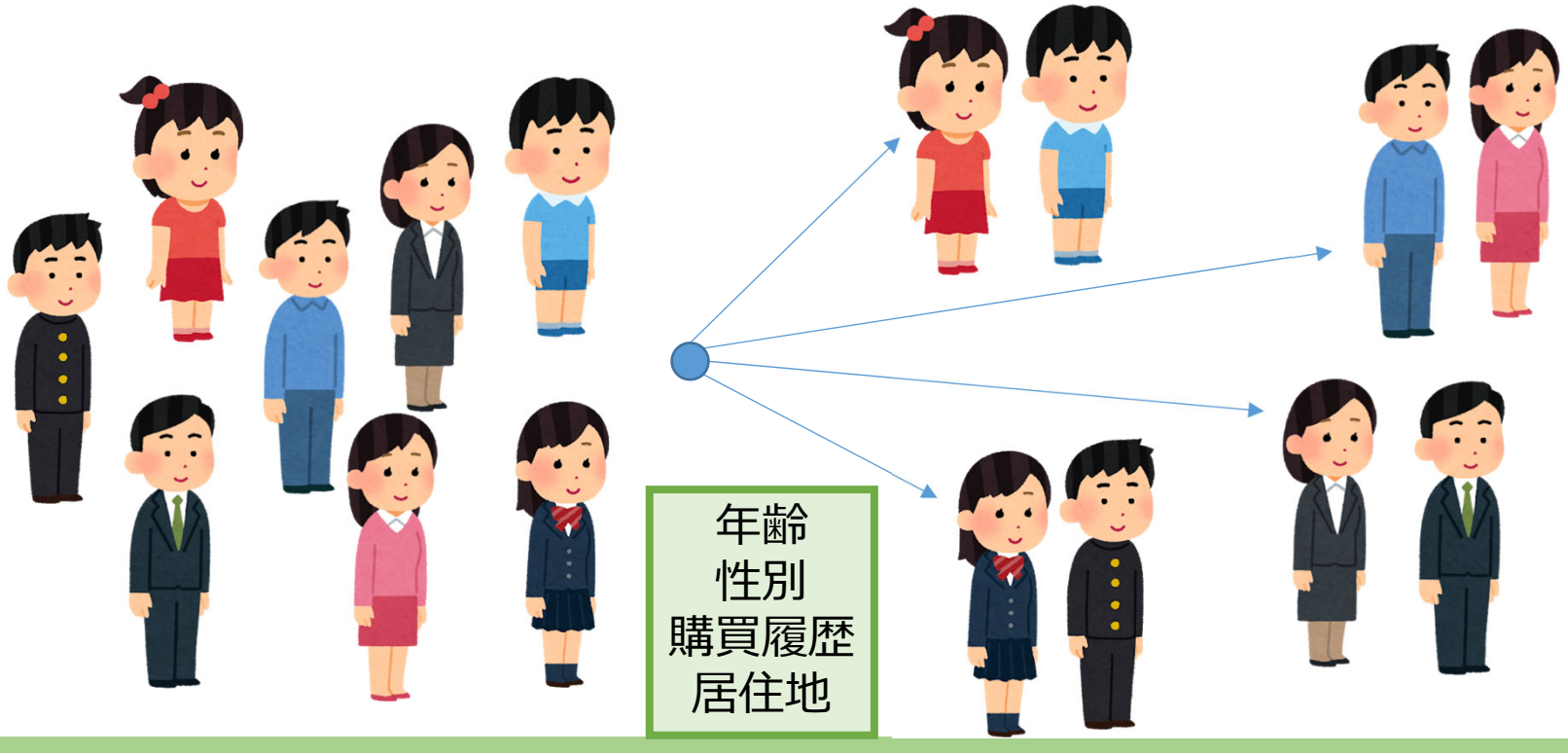
授業計画


Week 1: Introduction
Week 2: Basics
データの入手
データ加工処理
分析結果の可視化
可視化による見通し
決定木
相関ルール
クラスタ分析
回帰分析
ニューラル
ネットワーク
Week 2 & 3
Week 4 & 5
Week 6 & 7
モデル選択
クロスバリデーション
Week 8: 応用

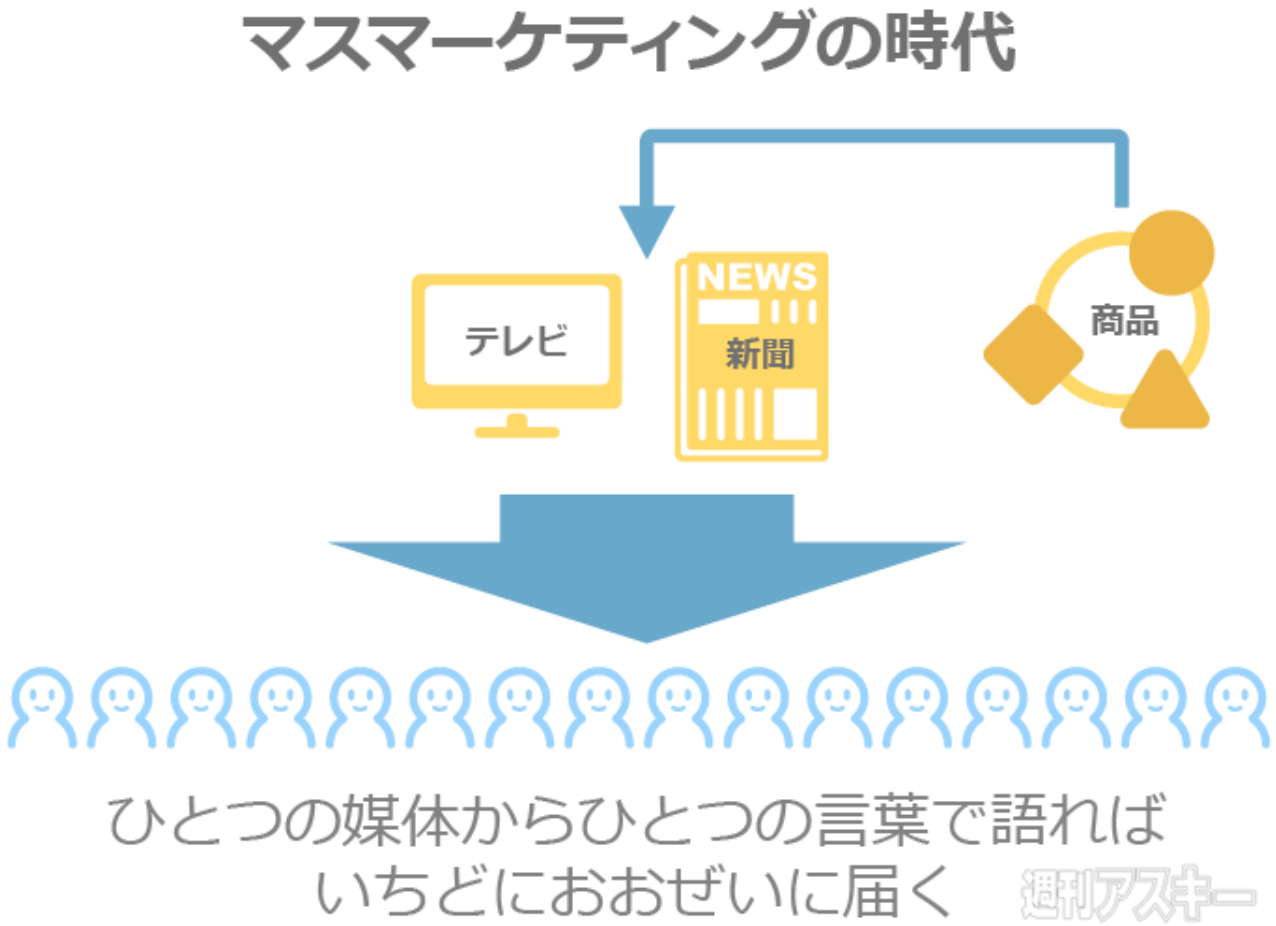
データマイニングの代表的な手法

(2) クラスター分析

似ているデータごとにデータをまとめて分類
→適切な商品を推奨できる



昔のマーケティング

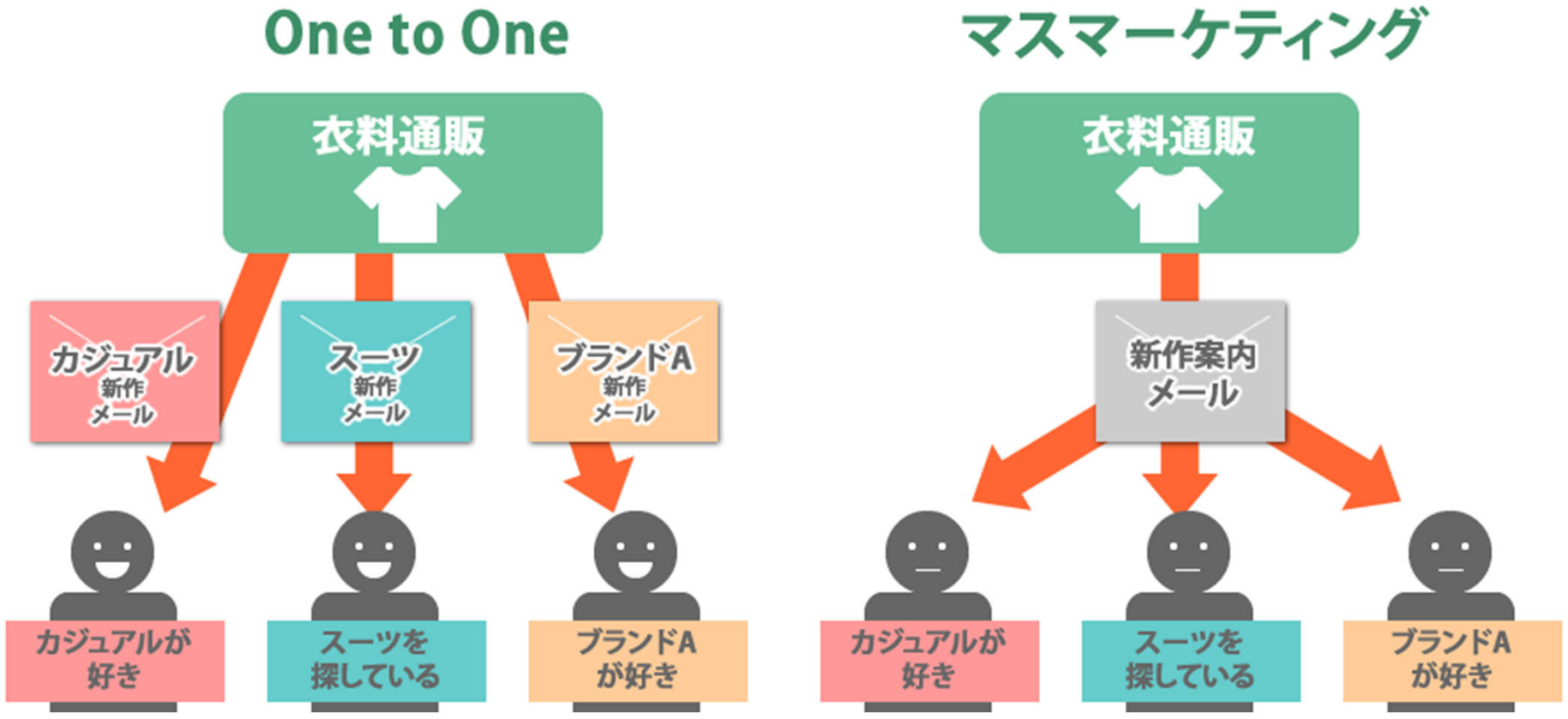


https://weekly.ascii.jp/elem/000/000/107/107931/

現代のマーケティング

https://weekly.ascii.jp/elem/000/000/107/107931/

マーケティング方法

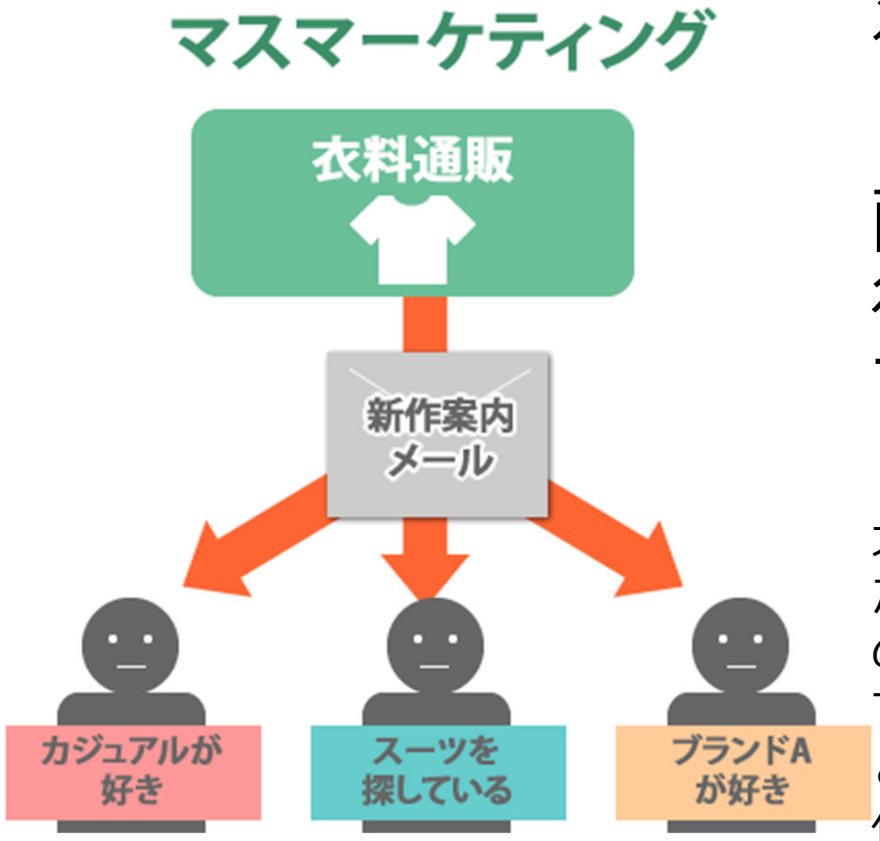


メルマガ
アプリでクーポン配布

TV CM
新聞広告

https://satori.marketing/marketing-blog/marketing-automation/onetoone_marketing/

マスマーケティング



対象を特定せず
（すべての消費者を対象）
画一化された方法を用いて
行うマーケティング戦略、
マーケティング活動。

大量生産と大量販売、マスメディアを用いた広告の大量投入を前提としており、市場の成長期にマーケットリーダー（ある市場で最大のシェアを持つ企業）が用いる手法としては有効だが、消費者の価値観が多様化した市場では特定のニーズに応えきれない場合がある

https://satori.marketing/marketing-blog/marketing-automation/onetoone_marketing/

One to oneマーケティング

一人一人の消費者のニーズや
購買履歴に合わせて、

個別に展開される
マーケティング活動

客の一人ひとりに違うクーポンを配信したり、異なったwebページを表示する施策をしたりすることで購買率があがります。

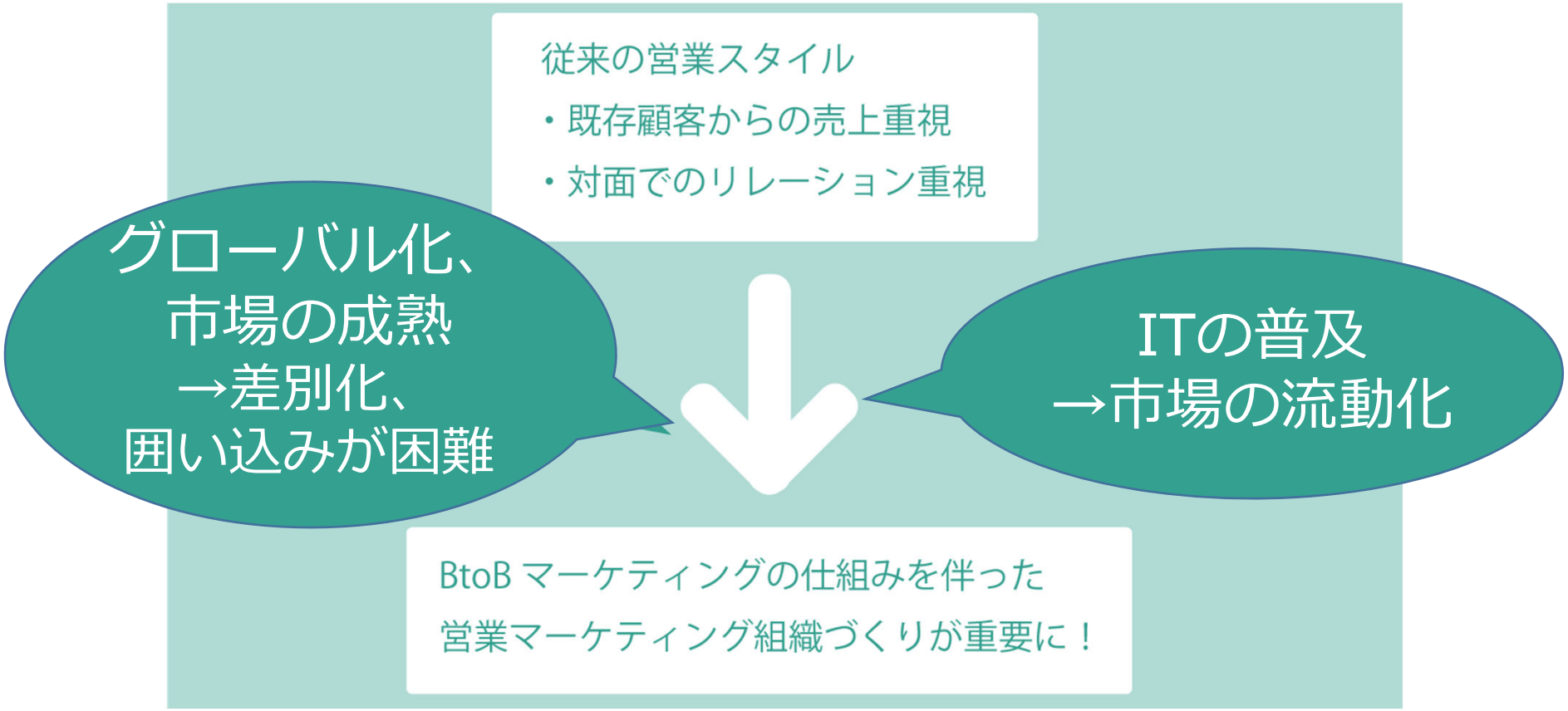
クラスター解析!

https://satori.marketing/marketing-blog/marketing-automation/onetoone_marketing/

B to B マーケティング

BtoB＝Business to Business

法人顧客相手（企業間取引）の取引



https://innova-jp.com/201610-btob-marketing/

One to oneマーケティング

広告費を10%以上削減しながら売上高は上昇
【すかいらーく】

2014年上半期の広告宣伝費
前年同期比で10%以上削減
売上高39億円2.9%の成長

年齢や性別、子供の有無などユーザーがアプリの登録時に設定した内容に合わせてクーポンを配信
例えば未成年のユーザーにお酒のクーポンを配信しても意味がないです。

ユーザーが行きたいと思い出す頃を見計らってクーポンを配信し

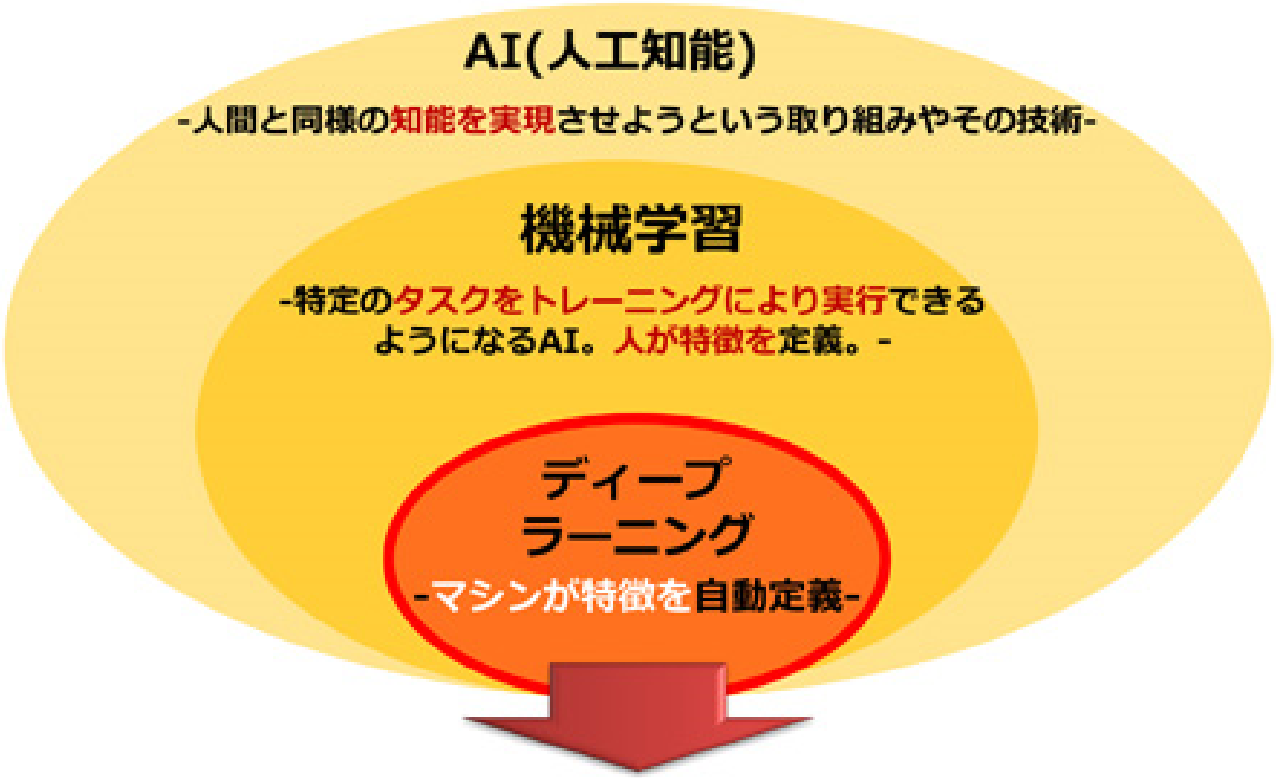
One to oneマーケティング

メリット：コストをかけず効果的なアプローチができる

- ユーザーが求めている情報を配信するので、広告でも**ユーザーにしつこいと思われない**
- カタログ印刷やDM印刷と違って**低コスト**
- 購買意欲の高いユーザーに適切なタイミングでアプローチができ、**購買につながる確率も高い**
- ITの活用により、**導入の手間やコストも少ない**

クラスター解析!

AIの三大分類



飛躍的なAI性能の向上
マーケティングの世界でも期待値急増
(第３次人工知能ブームへ)

応用例
- ◆ スパムメール検知機能
- ◆ 手書き文字認識
  (例) 郵便物の住所解釈
- ◆ 画像認識
  (例) 自動運転
- ◆ 音声認識
  (例) Siri, Alexa...

人が手作業でやってたら
大変な作業を
コンピューターで自動的に処理

https://markezine.jp/article/detail/29471

データマイニングと機械学習

機械学習とデータマイニングは交差する部分が大きく、
技法も同じなので混同されることが多い。
厳密に区別するのは難しい。

機械学習

訓練データから学んだ「既知」の特徴に基づく予測

既知の知識を再生成できるかどうかで性能を評価

→主に教師あり学習

データマイニング

それまで「未知」だったデータの特徴を発見

→教師なし学習のみ

教師なし学習 vs. 教師あり学習

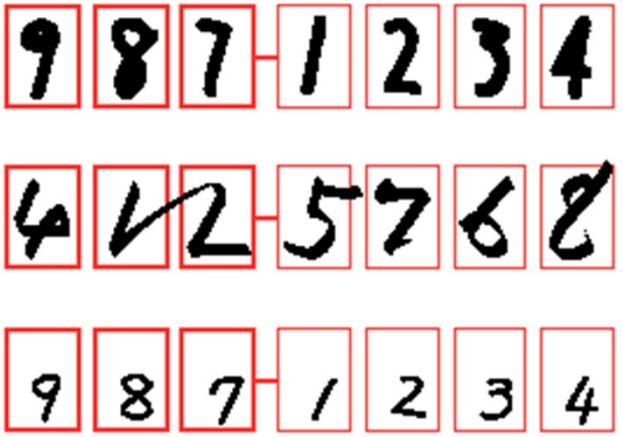
データマイニング　**例) 文字認識**

どんな文字があるか
推測する

答えがない

→どんな文字があるか、推測する

文字を解読する



答えがある

→どの文字に該当するか??考える
経験を積んで、精度を上げる

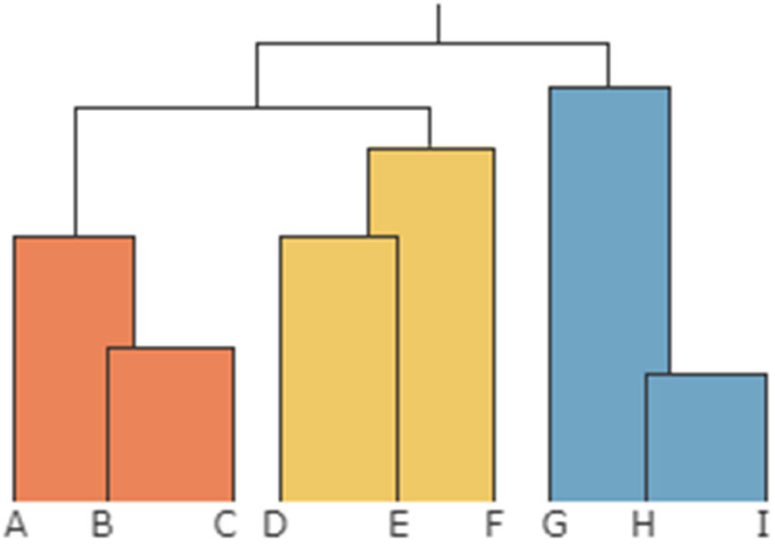
クラスター分析

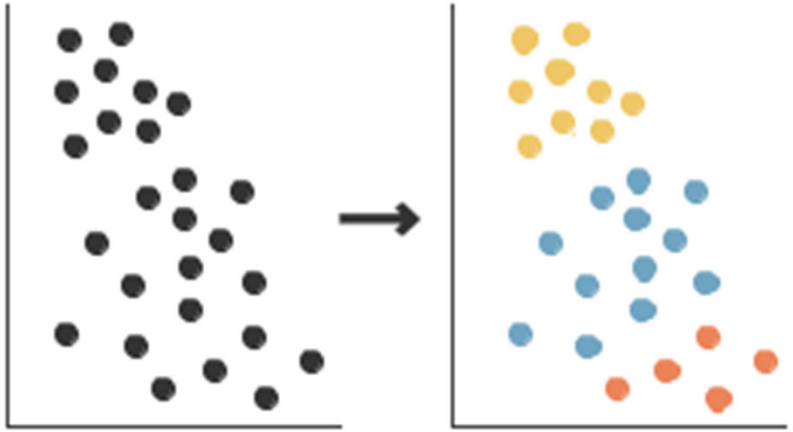
教師なし機械学習の一種
いくつのクラスターになるべきか、
といった答えはない。

クラスター=Cluster
→房、集団、群れ

階層型クラスター分析

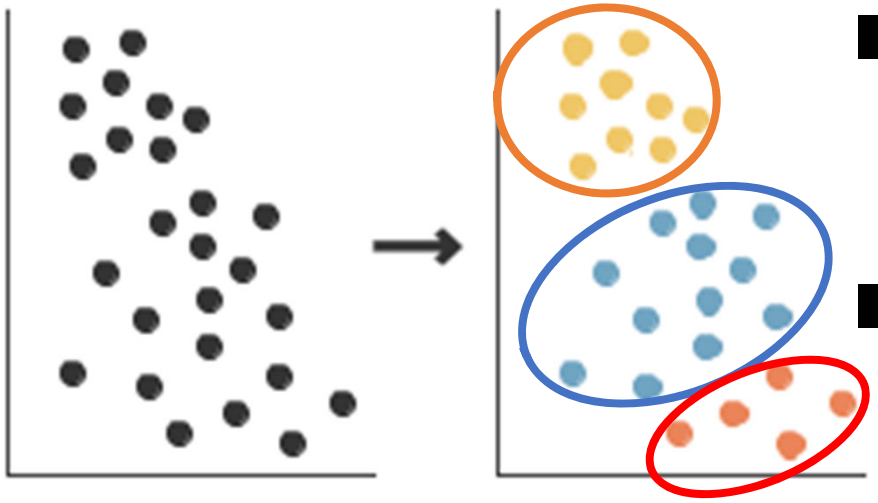
非階層型クラスター分析





https://promote.list-finder.jp/article/marke_all/cluster-analysis/

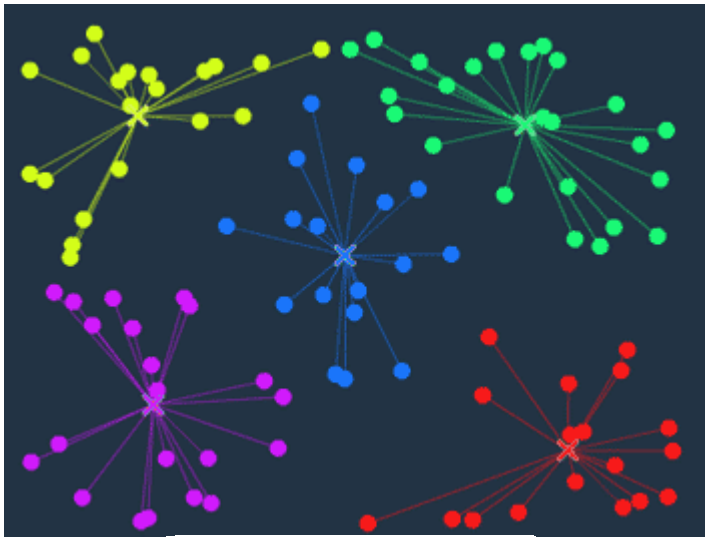
非階層型クラスター解析



- k-means法
- 混合ガウスモデル (Gaussian Mixture Model, GMM)
- EM アルゴリズム (Expectaion-Miximization Algorithm)

https://promote.list-finder.jp/article/marke_all/cluster-analysis/

k-means法



nstep=9

クラスタ数=5

step 0 クラスタの数を決める

step 1
各点にランダムにクラスタを割り当てる

step 2
クラスタの重心を計算

変化あり
2 に戻る

step 3
点のクラスタを、
一番近い重心のクラスタに変更する

クラスタの
組み換えなし

終了

分析手法の基礎：平均

	x (Math) [points]	y (50m run) [sec]
Harvey	90	8.8
Mike	100	7.6
Louis	98	9.5
Harold	87	10.2
Robert	80	12.4

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

N: number of samples

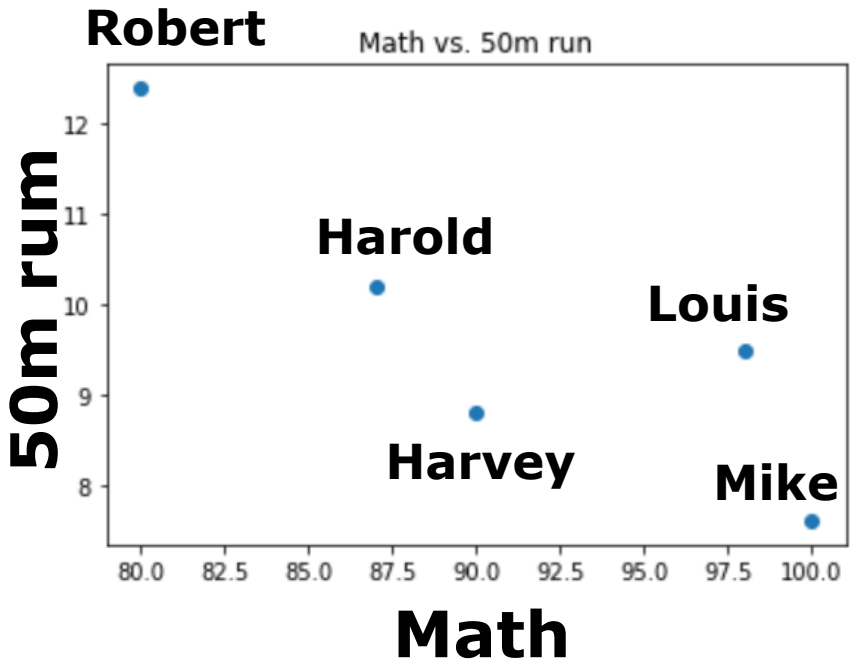
$$\bar{x} = \frac{90 + 100 + 98 + 87 + 80}{5} = 91$$

$$\bar{y} = \frac{8.8 + 7.6 + 9.5 + 10.2 + 12.4}{5} = 9.7$$

分析手法の基礎：平均

	x (Math) [points]	y (50m run) [sec]
Harvey	90	8.8
Mike	100	7.6
Louis	98	9.5
Harold	87	10.2
Robert	80	12.4

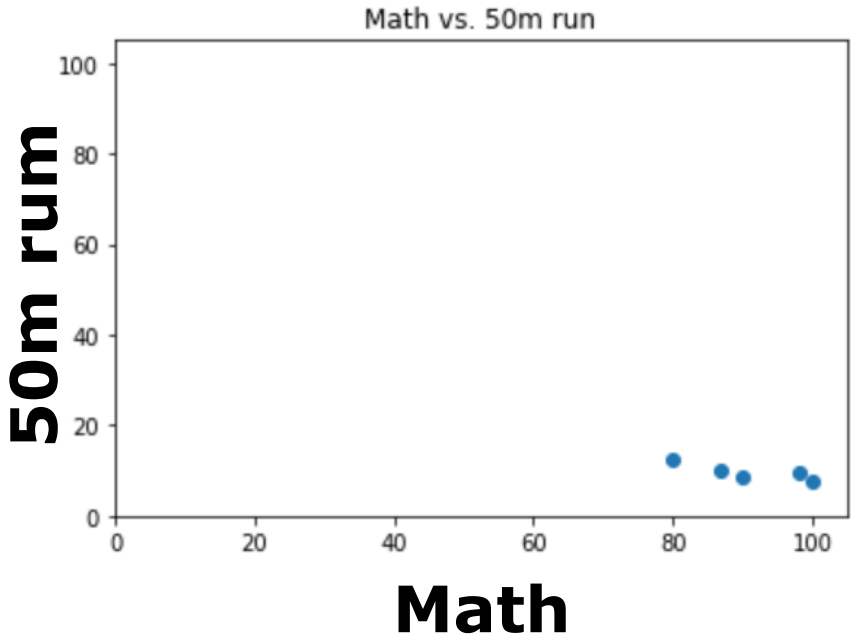
散布図でプロット



分析手法の基礎：平均

	x (Math) [points]	y (50m run) [sec]
Harvey	90	8.8
Mike	100	7.6
Louis	98	9.5
Harold	87	10.2
Robert	80	12.4

散布図でプロット



実際の数値で解析しようと思うと、
数学の点のほうが値が大きいので、
数学の点の影響が大きくなる。

標準が０、標準偏差が1になるように、データを規格化する。

分析手法の基礎：分散

	x (Math) [points]	y (50m run) [sec]
Harvey	90	8.8
Mike	100	7.6
Louis	98	9.5
Harold	87	10.2
Robert	80	12.4

σ^2：分散
σ：標準偏差

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

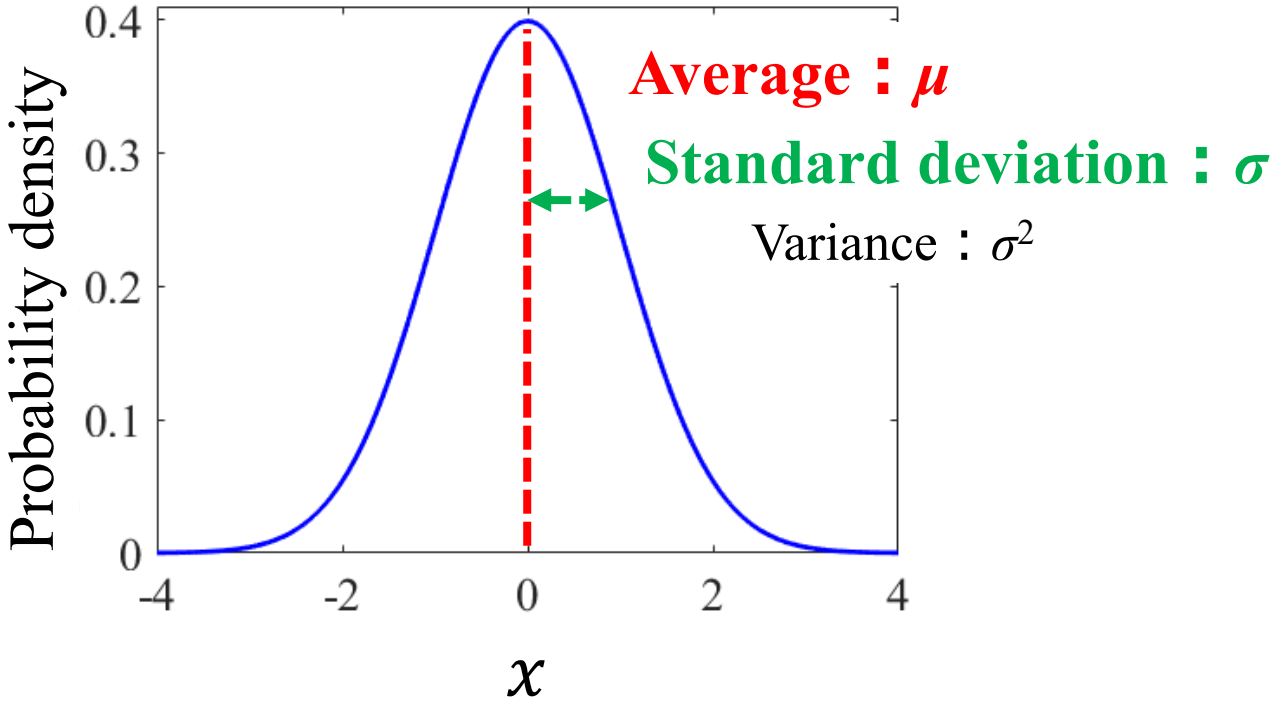
N: number of samples

$$\sigma_x^2 = \frac{(90-91)^2 + (100-91)^2 + (98-91)^2 + (87-91)^2 + (80-91)^2}{5} = 67$$

正規分布(ガウス分布)

More data around the average

Less data at the tail



Probability density function
確率密度関数

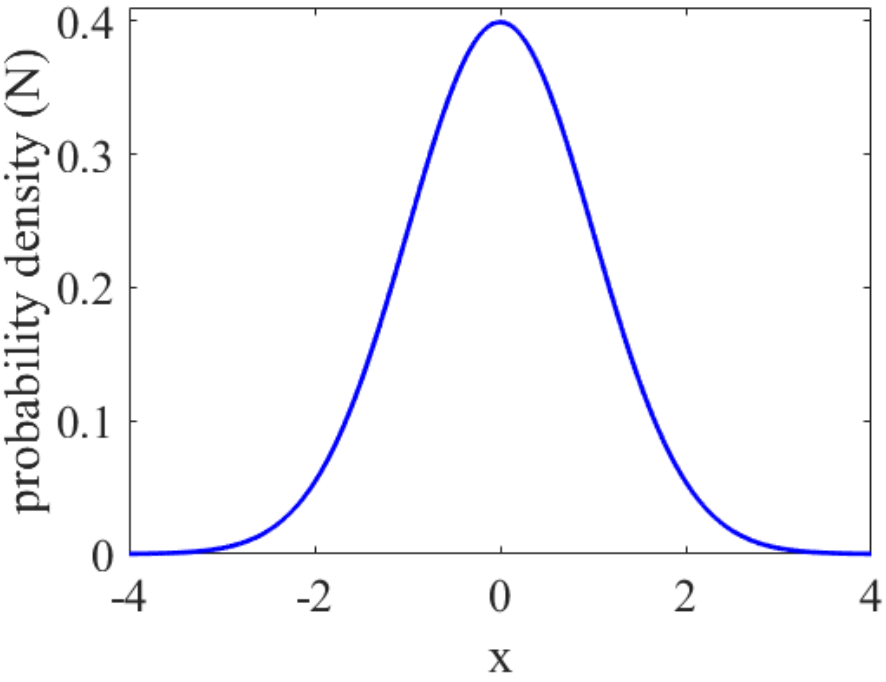
$$N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

標準正規分布

平均　$\mu = 0$

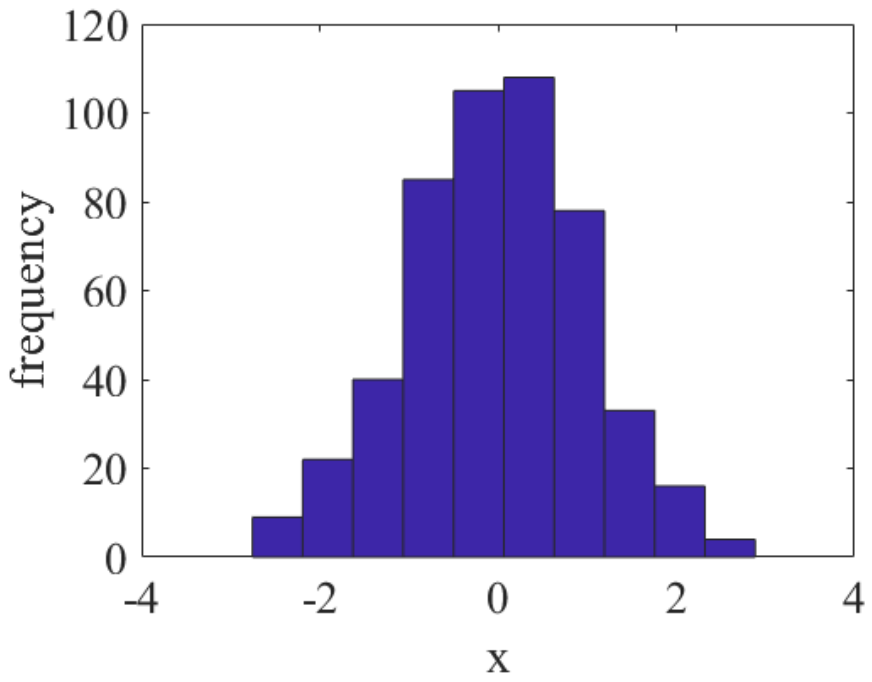
分散　$\sigma = 1$

$$N(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$



Probability density function

確率密度関数

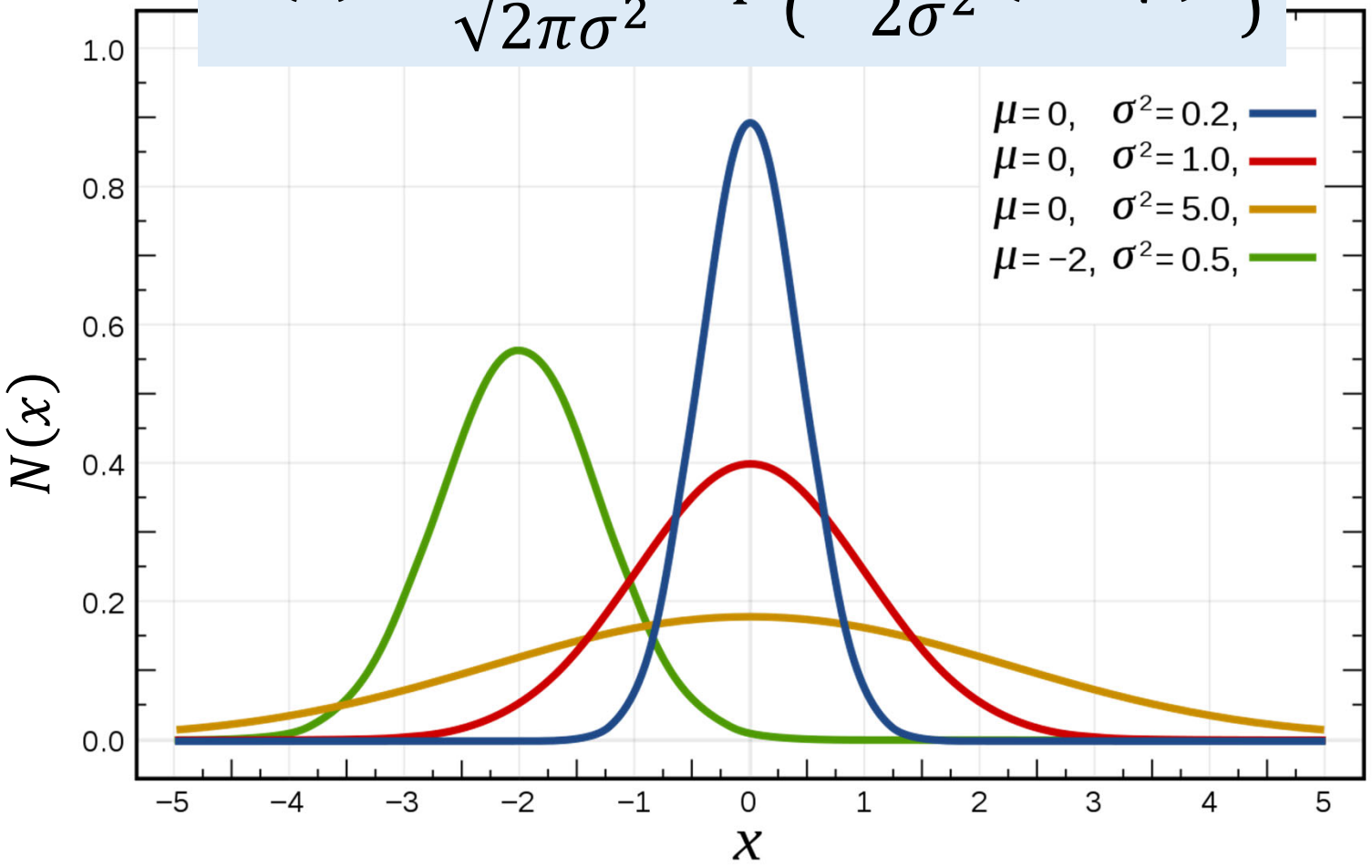


Histogram

ヒストグラム

正規分布(ガウス分布)

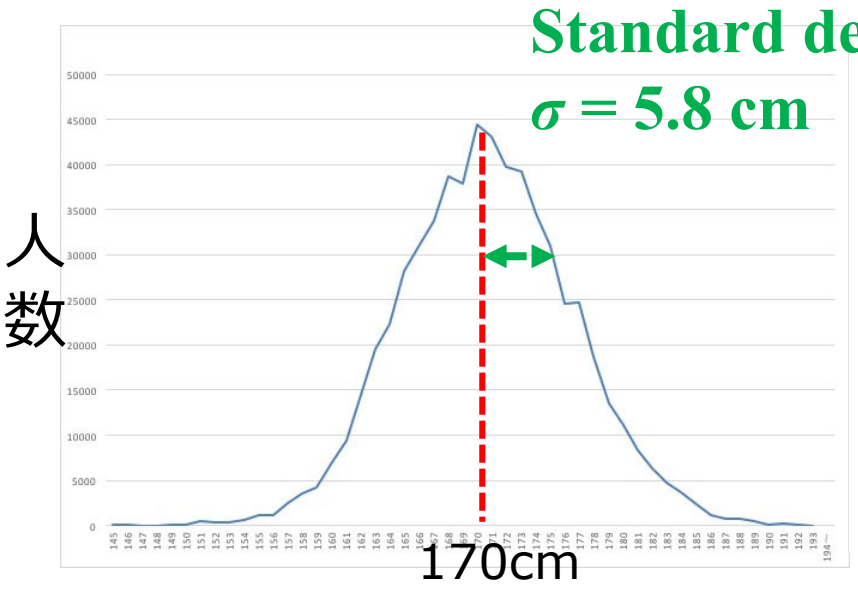
$$N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



身長は正規分布に従う

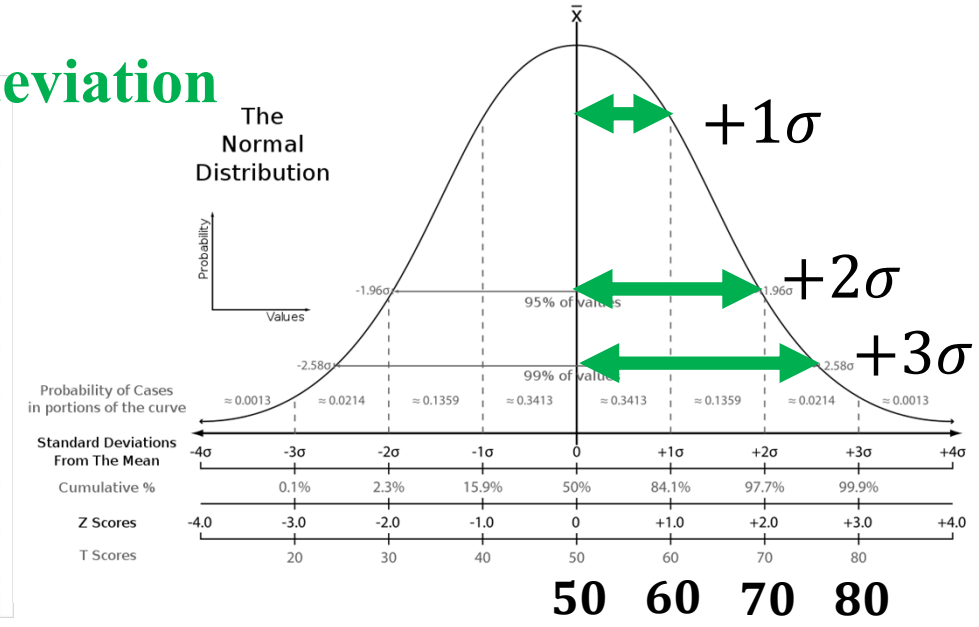
Distribution of height of 17-year-old boy

１７歳男子の身長の分布



Distribution of grades

成績の分布



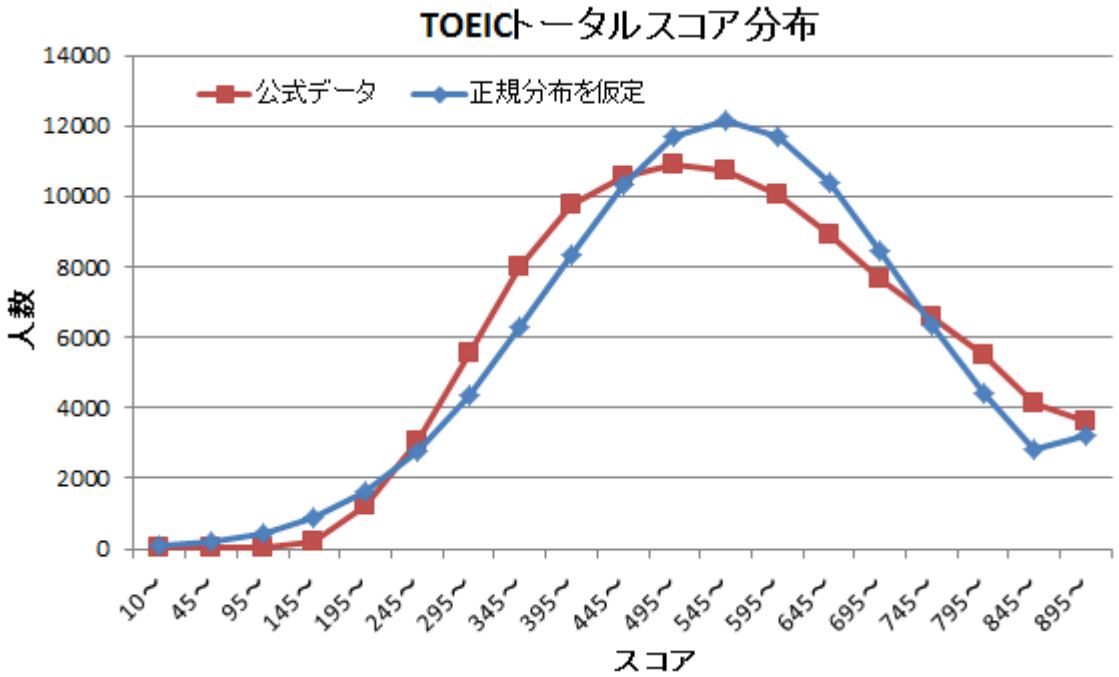
https://ai-trend.jp/basic-study/normal-distribution/example/

https://ja.wikipedia.org/wiki/%E5%81%8F%E5%B7%AE%E5%80%A4

成績は正規分布に従う(?)

個人の得点: x
平均点: $\bar{x}$
偏差値: y

$$y = \frac{x - \bar{x}}{\sigma} \cdot 10 + 50$$



- 母集団の数が十分大きいとき
- 平均点が高すぎないとき

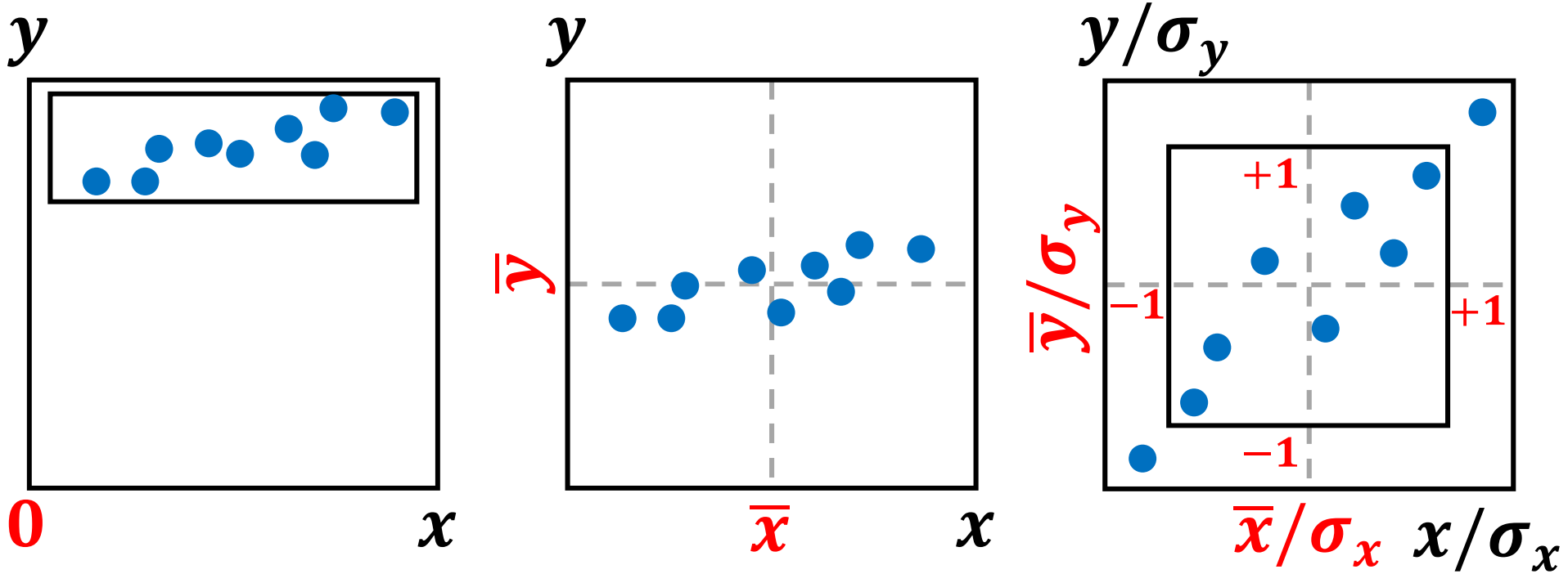
最高点は満点なので、
厳密には正規分布に
ならないでしょう。

大学の模試とか、複数科目の合計で、オール満点とかなかなか出にくかったら、正規分布に近づくかも。

https://sudillap.hatenablog.com/entry/2013/04/12/223031

分析手法の基礎:データの標準化

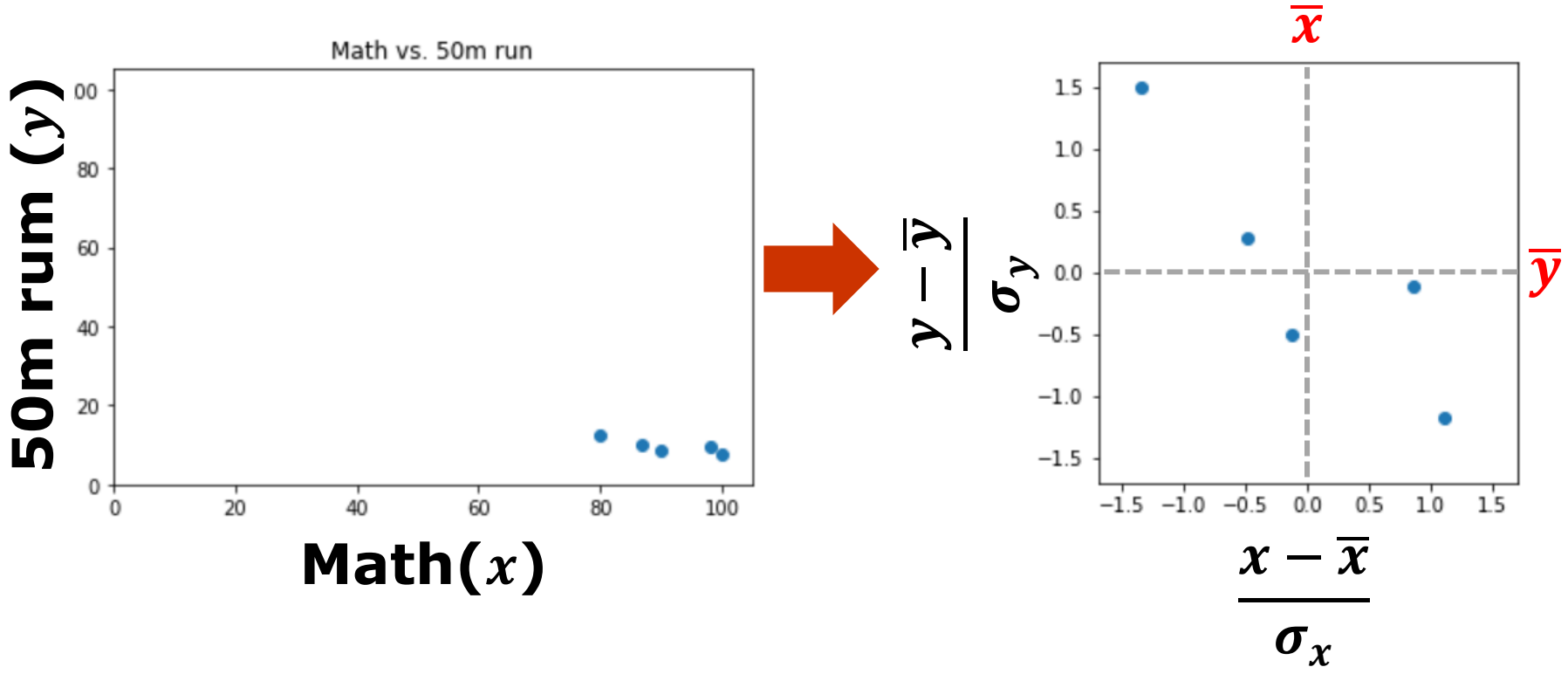
平均0、分散1になるようにデータを規格化することで、
複数の変数を評価しやすくする。



元のデータ → **平均を0にずらす** → **標準偏差が1になるように割る**

分析手法の基礎:データの標準化

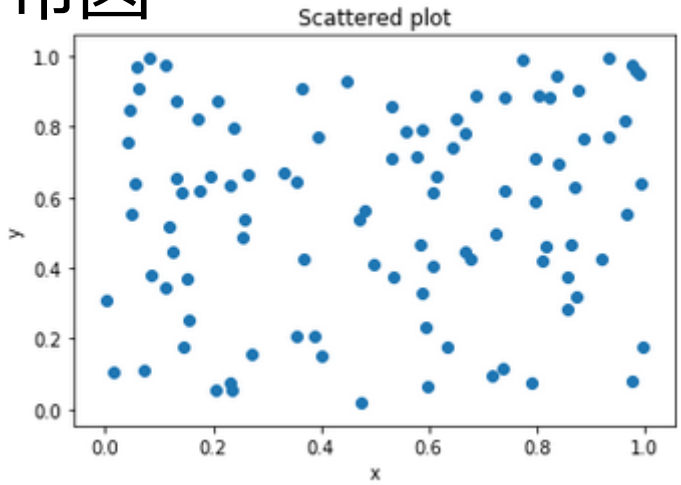
標準化は自分でもできますが、Scikit-Learnのパッケージを使うこともできます。



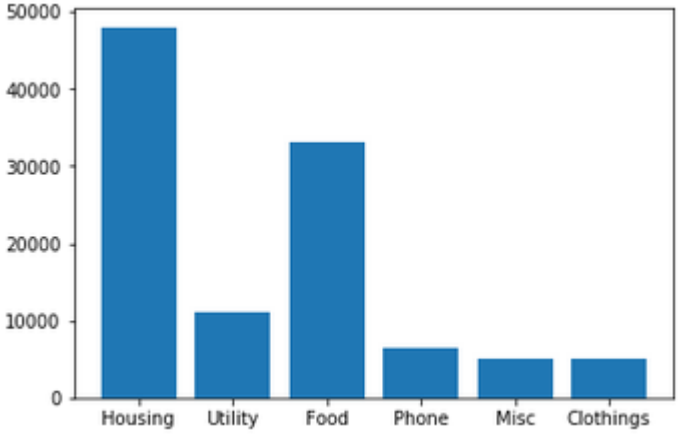
このデータに対して、
クラスタリング処理を行う。

データの可視化

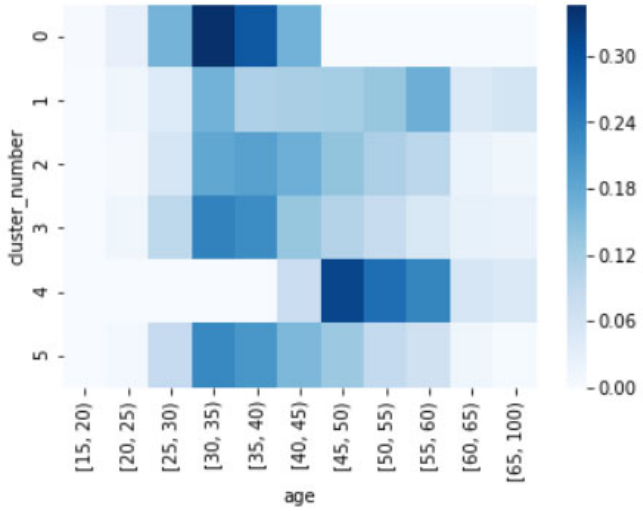
- 散布図



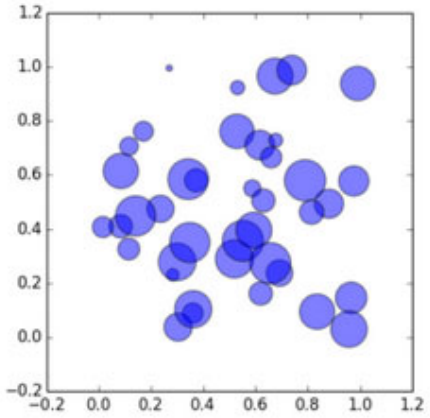
- 棒グラフ



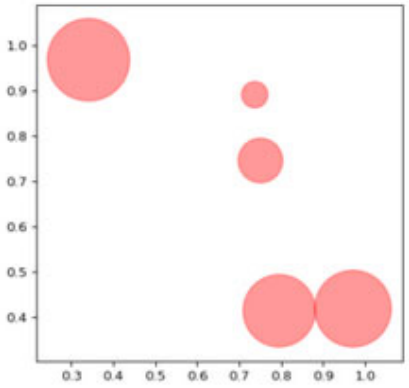
- ヒートマップ



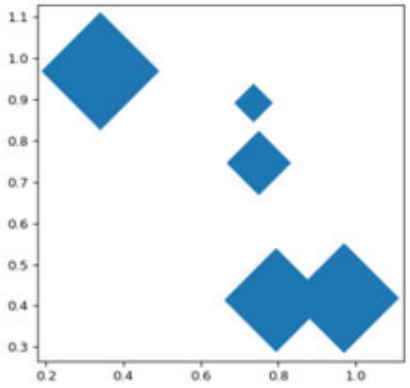
Python-Graph-Gallery.com



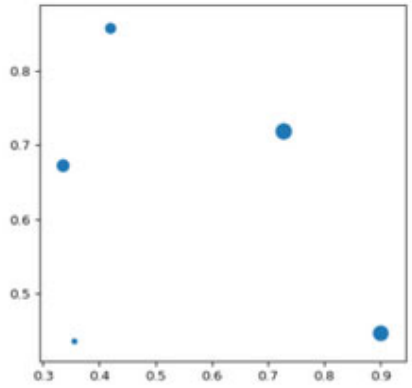
#270 Basic Bubble plot



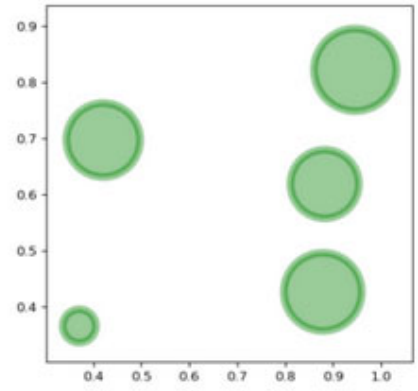
#271 Color of bubble



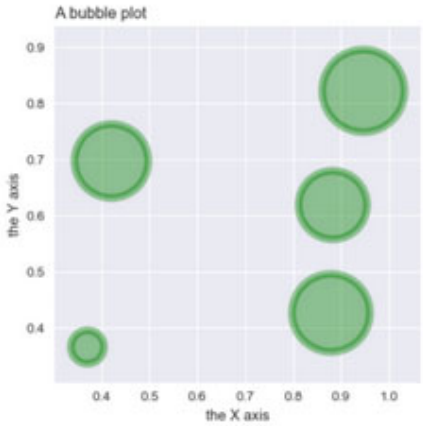
#271 Shapes of bubble plot



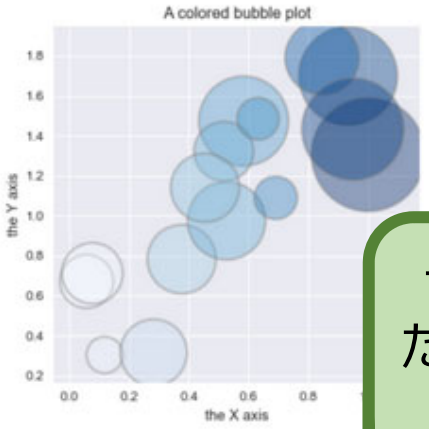
#271 Average size of bubble



#271 Edges of bubble plot



#271 Bubble with seaborn style



#272 Map a color to bubble

サンプルコードが
たくさんあるので、
参考になります!!

https://python-graph-gallery.com/bubble-plot/